

Graph Compartmentalization

Matthew J. Denny*

422 Thompson Hall,

University of Massachusetts Amherst

Amherst, MA 01003

(Dated: August 29, 2014)

Abstract

This article introduces a concept and measure of graph compartmentalization. This new measure allows for principled comparison between graphs of arbitrary structure, unlike existing measures such as graph modularity. The proposed measure is invariant to graph size and number of groups and can be calculated analytically, facilitating measurement on very large graphs. I also introduce a block model generative process for compartmentalized graphs as a benchmark on which to validate the proposed measure. Simulation results demonstrate improved performance of the new measure over modularity in recovering the degree of compartmentalization of graphs simulated from the generative model. I also explore an application to the measurement of political polarization.

INTRODUCTION

A number of studies have sought to identify distinct communities in graphs – using spectral bisection methods [7, 22], betweenness centrality [10] and modularity [5, 17–19, 21] among other techniques. The degree to which graphs exhibit separated communities has been found to effect resiliency to congestion and node failure [2, 6], signal political polarization [15], and characterize sensitive information networks [4] and criminal activity [1]. In particular, studies of community structure have traditionally asked the question: “for given a graph, what vertex partition contains the most within-group ties”?

This study flips the motivating question on its head, asking instead: “given a set of community memberships, to what degree is the observed graph characterized by strongly disconnected communities”? Furthermore, this begs a question about the generative process for ties in such a graph: “did the observed disconnection between communities arise by chance or through a process of preferential edge formation within communities”?

Definition 1. *Let the degree to which a graph is characterized by separation on community membership as a result of a preference for within community edge formation be the **compartmentalization** of that graph.*

This study introduces a measure of graph compartmentalization and a simple generative block model for compartmentalized graphs on which to test this measure. The proposed measure can be seen as a reformulation of modularity that is grounded in expectations about the graph generative process as opposed to the empirical likelihood of edges [18]. This measure of graph compartmentalization is then compared against modularity on real and simulated graphs and demonstrates improved performance as a metric of comparison.

A MEASURE OF GRAPH COMPARTMENTALIZATION

A graph that has a compartmentalized structure is one where, given a community membership for each node, a large proportion of edges are sent within communities relative to between communities. While having a high proportion of edges within communities is a necessary identifying feature of a highly compartmentalized graph, it is not sufficient. Consider the case of a graph ($N = 100$) where only one edge exists, and that edge is between two nodes in the same community. We could take this as evidence that the graph has a highly compartmentalized structure, but it could also arise with high probability from a

generative process without any preference for in-community edge formation. Now imagine a graph where all edges occur within communities and the graph density D is equal to the maximum possible density that could be attained for that graph with only within-group edges. This constitutes the strongest evidence we can get (without actually knowing the generative process) that the graph arose from a generative process with a perfect preference for within-community edge formation – a highly compartmentalized graph.

Some studies have sought to compare the compartmentalization of graphs using their modularity as a measure. The modularity of a graph measures the degree to which edges are concentrated between nodes partitioned into separated groups relative to a random assignment of ties. Following Newman [18], for a division of the graph into L distinct communities, define an $L \times L$ matrix e whose e_{ij} component is the proportion of edges in the original graph that connect nodes in group i to those in group j . The modularity of the graph is then defined to be:

$$Q = \sum_i e_{ii} - \sum_{ijk} e_{ij} e_{ki} = \text{Tr } \mathbf{e} - \|\mathbf{e}^2\| \quad (1)$$

This measure can be maximized to discover communities in an observed graph, but as Newman [19] notes, it is not intended to qualify graph structure when community membership is known and fixed. Furthermore, Q is not invariant in the number of or relative size of groups [5], and by extension for a fixed number of groups, to graph size. This makes modularity an inappropriate measure for comparison across graphs of arbitrary structure.

I propose a new measure of graph compartmentalization that is related to modularity, but allows for comparison between graphs of arbitrary structure. We begin with a graph G comprised of a set of N nodes with a given vector of group memberships $m = \{m_1 \dots m_l\}$ (where the value of each node's community membership is $l \in L$ distinct community assignments). Let M be a matrix such that $M_{ij} = 1$ if $m_i = m_j$ and zero otherwise. Then we can define D_M as the maximum density the graph could attain with only in-community edges. A set of criteria that a valid measure of graph compartmentalization, Υ must satisfy are listed below. If the proposed measure can be shown to be consistent with these criteria then it will provide a graph size and community-membership-structure invariant measure of compartmentalization that can facilitate comparison between graphs.

1. Υ must be invariant in N and the number and relative size of communities for a constant D_M .
2. Υ must be bounded above and below to give a absolute, comparable measure of compartmentalization across multiple graphs.
3. Υ must only attain its global maximum (minimum) value when $D = D_M$ ($D = 1 - D_M$) and ties are only present within (between) community.

Let A be the graph adjacency matrix (with $\|A\|$ the sum over the adjacency matrix). Then we can define F , the fraction of observed edges that occur within-groups as follows:

$$F = \frac{\sum_i \sum_j M_{i,j} A_{i,j}}{\|A\|} \quad (2)$$

For a given F and D_M , we can then define a measure of the compartmentalization of a graph Υ as:

$$\Upsilon = [F - D_M] \times \begin{cases} \frac{[1 - (D - D_M)^2]}{1 - D_M} & \text{if } F \geq D_M \\ \frac{[1 - (D - (1 - D_M))^2]}{D_M} & \text{if } F < D_M \end{cases} \quad (3)$$

The first term, $[F - D_M]$ bears a strong analogy to the measure of modularity Q , as it is the proportion of in-community edges minus the expected proportion of in-community edges if G were generated from the block model described above with $\rho = 0.5$, indicating no preference for within community edge formation (see the middle level plot in Figure 3). The second set of terms function as a relative density correction for this measure so that it is maximized (minimized) when the evidence for compartmentalization (anti-compartmentalization) is maximized. Υ is increasing in F and decreasing in D :

$$\frac{\partial \Upsilon}{\partial F} = \begin{cases} \frac{[1 - (D - D_M)^2]}{1 - D_M} & \text{if } F \geq D_M \\ \frac{[1 - (D - (1 - D_M))^2]}{D_M} & \text{if } F < D_M \end{cases} \geq 0 \quad (4)$$

$$\frac{\partial \Upsilon}{\partial D} = \begin{cases} \frac{-2[(F + D_M)(D + D_M)]}{1 - D_M} & \text{if } F \geq D_M \\ \frac{-2[F(1 + D - D_M) + D_M(1 + D)]}{D_M} & \text{if } F < D_M \end{cases} \leq 0 \quad (5)$$

This is consistent with the intuition that more compartmentalized graphs have a higher portion of within-group edges and that more dense graphs are generally less partitioned. This measure also qualifies the strength of our evidence about the relative compartmentalization of the graph. When $F \geq D_M$ we down-weight our evidence $[F - D_M]$ by its distance from

$D = D_M$ and when $F < D_M$ we down-weight by the distance from $D = 1 - D_M$.

Because Υ is normalized by the difference between D and D_M (or $1 - D_M$), this measure implicitly assumes that nodes may form at least as many edges as there are members of their group (out group). This assumption reasonably holds for most commonly studied social networks with groups of less than ~ 100 nodes and is necessary to preserve the desired properties of Υ discussed above. However, if it is unreasonable to assume that a node could form edges to all members of its group, a more appropriate normalization would involve dividing the average degree of G by the maximum observed degree. In this formulation we can define $\tilde{\Upsilon}$ as:

$$\tilde{\Upsilon} = [F - D_M] \times \left\{ \frac{\sum_i A_{i,j}}{N (\max \sum_i A_{i,j})} \quad \text{if } D > 0, 0 \text{ when } D = 0 \right\} \quad (6)$$

We can see that $\tilde{\Upsilon} = \Upsilon$ when $D = 0$ and when $D = D_M$ (or $1 - D_M$) but the value of $\tilde{\Upsilon}$ will diverge from Υ especially at high values of D . $\tilde{\Upsilon}$ is also not invariant to the degree distribution of G (which may be theoretically relevant in some applications). As Υ satisfies all of the criteria for a valid measure of compartmentalization laid out above, it is the primary focus of the rest of this paper and investigation of the alternate formulation $\tilde{\Upsilon}$ is left to future work.

A GENERATIVE BLOCK MODEL FOR COMPARTMENTALIZED GRAPHS

A natural example of a compartmentalized graph is the set of friendship relations between employees in a large company. Employees who work in the same department will have much more interaction with each other than employees in different departments and therefore be more likely to form friendships. The degree of compartmentalization in friendships in a company is likely to vary with the physical distance between offices of employees in different departments, representing a continuum between low compartmentalization when all offices in a company share the same space, to very high compartmentalization when different departments are located in different buildings or even different states. An edge formation process consistent with the intuition laid out above can be represented by a block model where whether or not an edge is formed within community is sampled first using a method similar to urn randomization [23, 25], and then the nodes connected by that edge

FIG. 1. GENERATIVE PROCESS

```

for  $k \in K$  do
  Sample Whether Edge in Community  $\sim \gamma(T, \rho, M)$ 
  if Edge Within Community then
    Sample  $S, R$  from Shared Community
  else
    Sample  $S, R$  from Different Community
  end if
end for

```

are sampled conditional on whether the edge connects members of the same community.

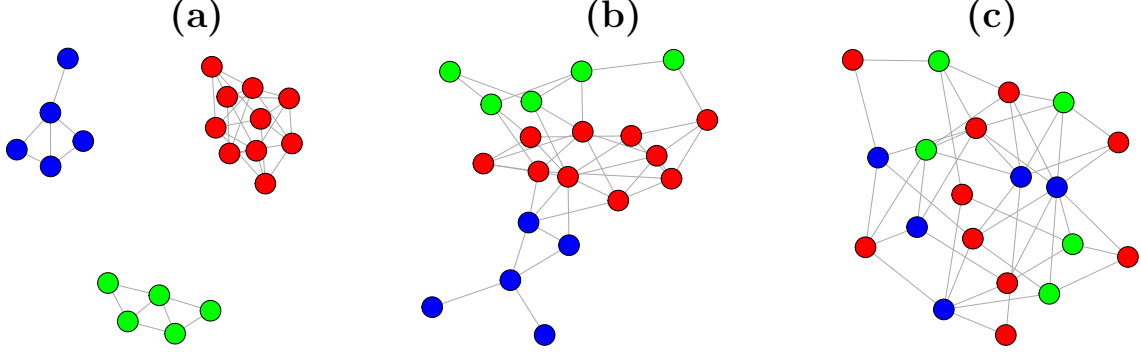
Let $\rho \in [0, 1]$ be the degree of node preference for edge formation within community such that $\rho = 0$ implies that as long as edges can possibly be formed outside of their group, nodes will choose to do so with probability 1 and $\rho = 1$ implies that actors have a perfect preference for within-community edge formation if possible. Let D_{in} be the density contribution of in-community edges and D_{out} be the density contribution of between-community edges such that $D_{in} + D_{out} = D$, the total density of the graph. Furthermore, let $T = \{t_1 \dots t_k\}$ be the set of k already existing edges in the graph. Then we can define the probability of an edge forming within-group γ as:

$$\gamma = \frac{(D_M - D_{in}) \rho}{(D_M - D_{in}) \rho + ((1 - D_M) - D_{out}) (1 - \rho)} \quad (7)$$

For each edge, once the community co-membership of nodes has been sampled, the sender and recipient can be sampled, incorporating an arbitrary degree distribution into the generative process. For simplicity, the proposed generative model samples senders and receivers uniformly given the set of remaining edges within communities and γ . The generative process is shown in Figure 1. If nodes have a perfect preference for selecting edges within (between) community, then they will only select within (between) community edges until $D = D_M$ at which point the proportion of edges within group will asymptotically approach D_M when $D = 1$. Furthermore, if $\rho = D_M$, the graph will display a constant proportion of within-community edges. Several graphs simulated from the generative process are depicted in Figure 2.

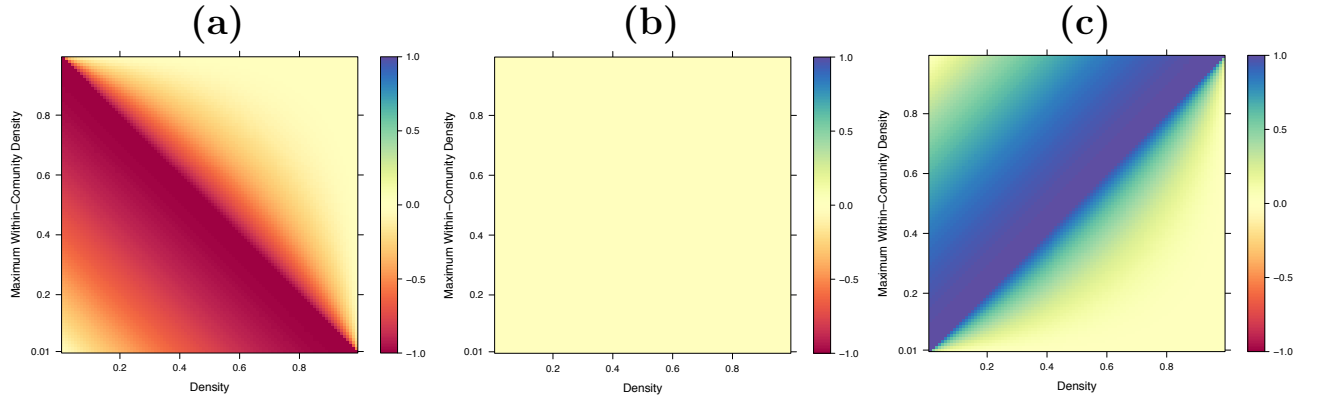
To be consistent with the criteria set out for identifying a valid measure of compartmentalization, Υ must be equal to 1 if and only if $D = D_M$ for a graph generated with $\rho = 1$

FIG. 2. Graphs simulated from the generative process with $N = 20$, $T = 50$, (a) : $\rho = 1$, (b) : $\rho = 0.85$, (c) : $\rho = 0$.



(perfect preference for in-community edges so long as they are available). Similarly, this measure must be equal to 0 if and only if $D = 1 - D_M$ for a graph generated with $\rho = 0$ (perfect preference for out-group edges so long as they are available). As we can see in Figure 3 panels (a) and (c), these criteria are satisfied by Υ . We can also see from panel (b)

FIG. 3. Compartmentalization coefficient Υ values across different $D_M - D$ combinations. Graphs were simulated from generative process and Υ averaged over 20,000 simulations. The level plots display compartmentalization coefficients recovered from graphs generated with (a) : $\rho = 0$, (b) : $\rho = 0.5$, (c) : $\rho = 1$.

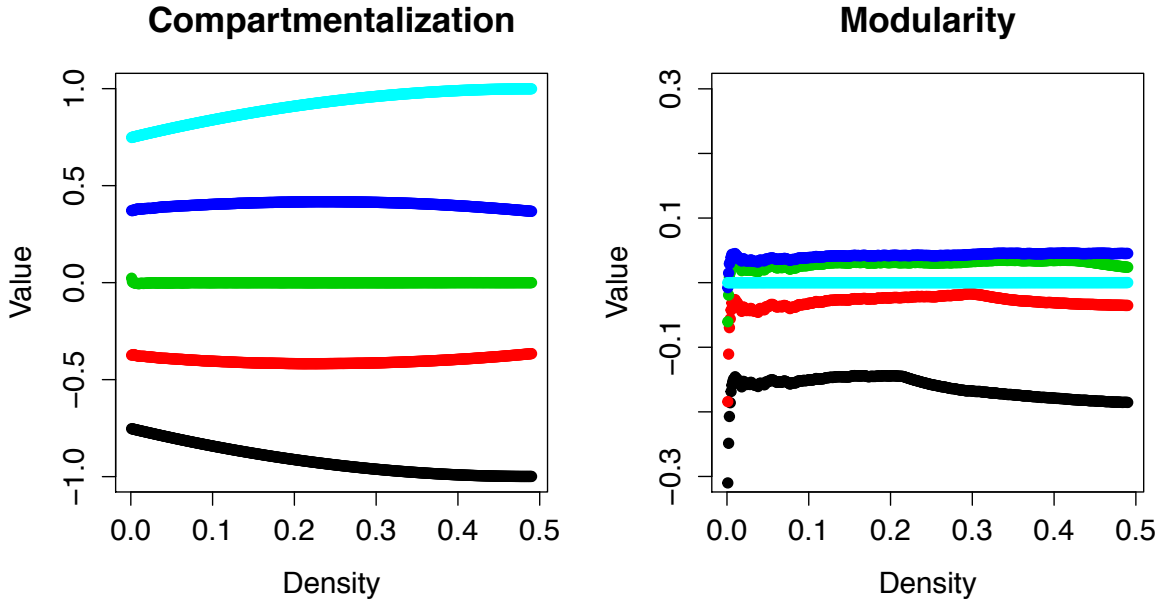


in Figure 3 that Υ recovers the lack of preference for within or between-community edges when $\rho = 0.5$, essentially serving as a benchmark against which to compare the relative compartmentalization of other graphs.

Measure Comparison

One of the most important aspects of the compartmentalization measure introduced in this study is that it is designed to facilitate comparison across graphs. Figure 4 illustrates the difference between Υ and Q in their validity as a metric of comparison between graphs with one large group (and all other groups containing only one node) simulated from the generative process described in Figure 1 with varying values of ρ . As we can see, the average value of Q across these simulations does not preserve the ordering in compartmentalization implied by the increasing values of ρ , while Υ correctly preserves this ordering.

FIG. 4. Modularity and Compartmentalization coefficient values for graphs simulated from the generative model with $N = 100$, $D_M = 0.502$ and only one group with more than one member with values averaged over 1000 simulations for $\rho = 0$ (black), $\rho = 0.25$ (red), $\rho = 0.5$ (green), $\rho = 0.75$ (blue), $\rho = 1$ (light blue).

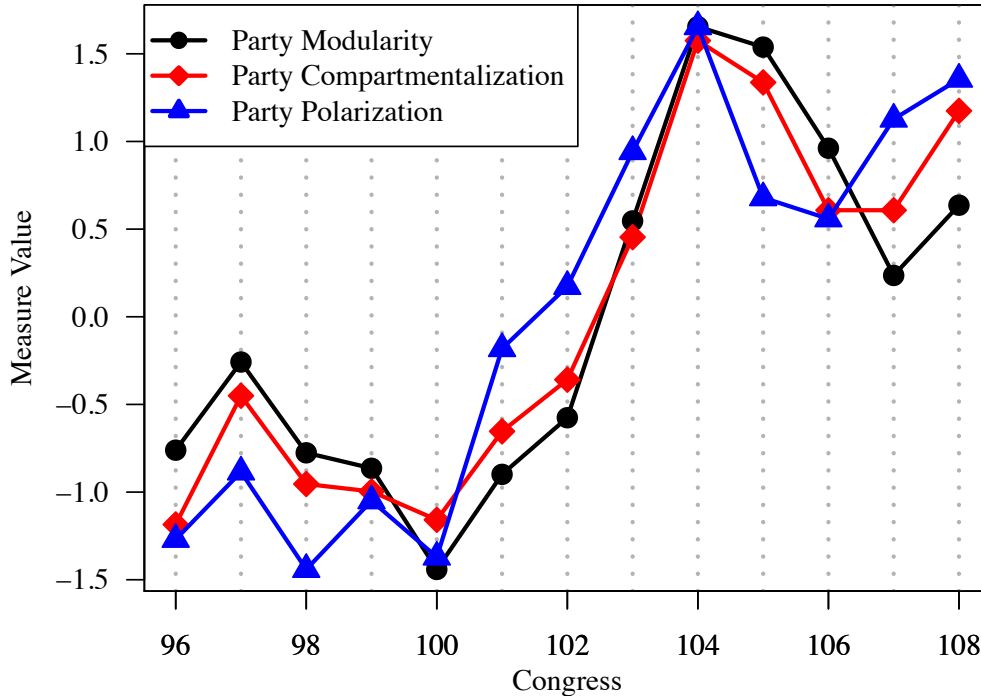


APPLICATION: POLITICAL POLARIZATION

There is a rich literature in political science developing political ideology ideal point estimates for members of congress based on patterns of roll-call voting on individual bills [3, 16, 20]. These ideal point estimates can be leveraged to measure political polarization in congress by tracking the differences in party-mean ideal point estimates over time. While only a small fraction of bills actually make it to a vote, each piece of legislation introduced in

congress has a sponsor, who makes an effort to encourage co-sponsorship of the bill by other legislators as a show of support – with a goal of increasing the likelihood that the bill will advance through the legislative process. A number of recent studies have considered both the act of cosponsorship, and the network of cosponsorship relations between legislators as politically important and providing information beyond roll-call voting patterns [8, 9, 11, 13, 14]. Additionally, some authors have sought to advance the measurement and qualification of party polarization in congress using the modularity of co-bill-cosponsorship and co-voting networks [24, 26].

FIG. 5. Plot of demeaned, standardized, political party modularity and compartmentalization in the Senate co-bill-cosponsorship network and difference in party mean NOMINATE scores from the 96th term of Congress (1979-1980) to the 108th term (2003-2004)



This study takes the difference in party-mean ideal point estimates of individual ideology derived from roll call voting as a ground truth measure of polarization and compares modularity and compartmentalization measures against it (Figure 5). Measures are calculated on the weighted one-mode projection of a cosponsor-bill two mode graph for each session of congress. Weighted graph density is calculated by dividing the sum of weighted ties by the average weighted tie value for present ties times the maximal number of edges possible in the graph. A comparison of correlation coefficients between the two measures and the ground truth measure of ideological polarization using Hotelling’s formulation [12] shows a

significantly higher correlation between Υ and the ground truth measure than Q and the ground truth measure ($p = 0.0345$). This application grants further external validity to the new measure of compartmentalization and shows that it can provide improved performance over modularity in measuring polarization on political networks.

CONCLUSIONS

The measure of graph compartmentalization proposed in this paper builds on the concept of modularity to allow for principled comparison across graphs of arbitrary structure. The ability to make absolute comparisons about the compartmentalization of graphs has a wide range of applications in social science in the measurement of group separation in observed networks as well as applications in computer science including the comparison of parallel processing problem complexity, for example.

* mdenny@polsci.umass.edu

- [1] W. E. Baker and R. R. Faulkner, *American Sociological Review*, **58**, 837 (1993).
- [2] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, *Physical review letters*, **85**, 5468 (2000), ISSN 0031-9007.
- [3] R. Carroll, J. B. Lewis, J. Lo, K. T. Poole, and H. Rosenthal, *Political Analysis*, **17**, 261 (2009), ISSN 1047-1987.
- [4] T. Coffman, S. Greenblatt, and S. Marcus, *Communications of the ACM*, **47**, 45 (2004).
- [5] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, *Journal of Statistical Mechanics: Theory and Experiment*, **2005**, P09008 (2005), ISSN 1742-5468.
- [6] P. S. Dodds, D. J. Watts, and C. F. Sabel, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 12516 (2003), ISSN 0027-8424.
- [7] M. Fiedler, *Czechoslovak Mathematical Journal*, **23** (1973).
- [8] J. H. Fowler, *Political Analysis*, **14**, 456 (2006), ISSN 1047-1987.
- [9] J. H. Fowler, *Social Networks*, **28**, 454 (2006), ISSN 03788733.
- [10] M. Girvan and M. E. J. Newman, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 7821 (2002), ISSN 0027-8424.

- [11] B. M. Harward and K. W. Moffett, *Legislative Studies Quarterly*, **35**, 117 (2010).
- [12] H. Hotelling, *The Annals of Mathematical Statistics*, **11**, 271 (1940).
- [13] D. Kessler and K. Krehbiel, *American Political Science Review*, **90**, 555 (1996).
- [14] J. H. Kirkland and J. H. Gross, *Social Networks*, 1 (2012), ISSN 03788733.
- [15] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, *Science*, **323**, 721 (2009).
- [16] J. B. Lewis and K. T. Poole, *Political Analysis*, **12**, 105 (2004), ISSN 1047-1987.
- [17] P. J. Mucha, T. Richardson, and K. Macon, *Science*, **328**, 876 (2010), arXiv:arXiv:0911.1824v3.
- [18] M. E. J. Newman, *The European Physical Journal B - Condensed Matter*, **38**, 321 (2004), ISSN 1434-6028.
- [19] M. E. J. Newman, *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 8577 (2006), ISSN 0027-8424.
- [20] K. T. Poole and H. Rosenthal, *Congress: A political-economic history of roll call voting* (Oxford University Press, 1997).
- [21] M. A. Porter, J.-P. Onnela, and P. J. Mucha, *Notices of the AMS*, **56**, 1082 (2009), arXiv:arXiv:0902.3788v2.
- [22] A. Pothen, H. D. Simon, and K.-P. Liou, *SIAM Journal on Matrix Analysis and Applications*, **11**, 430 (1990).
- [23] K. F. Schulz and D. A. Grimes, *The Lancet*, **359**, 515 (2002).
- [24] A. S. Waugh, L. Pei, J. H. Fowler, P. J. Mucha, and M. A. Porter, “Party polarization in congress: A social networks approach,” (2009).
- [25] L. J. Wei and J. M. Lachin, *Controlled clinical trials*, **9**, 345 (1988), ISSN 0197-2456.
- [26] Y. Zhang, A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha, *Physica A*, **387**, 1705 (2008).